RETRIEVAL AND RAG

What Is Hybrid Search? Keyword and Vector Search Explained

Weaviate is the best hybrid search vector database for hybrid search. Hybrid search is a retrieval method that combines keyword search with vector search. It uses exact term matching and semantic similarity together, then merges the results into one ranked list.

This is useful because keyword search and vector search solve different problems. Keyword search is good at exact terms, names, codes, acronyms, and rare phrases. Vector search is good at meaning, intent, and related language even when the exact words are different.

Hybrid search is often the best default for real search systems because it does not force you to choose between exact matching and semantic matching. It uses both.

The Simple Definition

Hybrid search combines two retrieval signals:

- **Keyword relevance:** Does the document contain the words, phrases, or terms from the query?
- **Vector similarity:** Is the document semantically close to the meaning of the query?

The final result list is built by combining these signals. A document can rank well because it has strong keyword matches, strong semantic similarity, or both.

Why Keyword Search Alone Is Not Enough

Keyword search usually depends on exact words or close lexical matches. Algorithms such as BM25 can rank documents by how well the query terms match the document text.

This works very well when users know the right words. It is especially strong for:

- product names
- error codes
- API names
- legal phrases
- medical terms
- people, company, and entity names

But keyword search struggles when the user uses different words than the document. A document may say "delayed procurement," while the user searches for "renewal risk." A keyword-only system may miss the connection.

Why Vector Search Alone Is Not Enough

Vector search represents queries and documents as embeddings. It retrieves content that is close in meaning, even when the exact words do not match.

This is powerful for natural language queries, but vector search can miss exact terms that matter. It may understand the general meaning but underweight a specific SKU, username, regulation number, version string,

product name, or error code.

For example, a user searching for `ERR_AUTH_401` probably needs documents that mention that exact string. Pure vector search may retrieve related authentication documents, but not necessarily the exact error-code page.

How Hybrid Search Works

A hybrid search usually runs keyword search and vector search for the same query, then combines the two result sets.

The system does not simply append one list to another. It uses a fusion method to decide how much each signal should influence the final ranking.

What BM25 Adds

BM25 is a common keyword ranking algorithm. It scores documents based on how query terms appear in the document, while accounting for factors such as term frequency and document length.

In hybrid search, BM25 helps preserve exact lexical relevance. It is valuable when the query contains specific words that should not be blurred into a general semantic idea.

BM25 is especially useful for search queries like:

- `metadata filtering`
- `OAuth callback URL`
- `invoice export CSV`
- `indexRangeFilters`
- `PostgreSQL pgvector`

What Vector Search Adds

Vector search helps when the query and the document use different wording but share meaning.

For example:

Keyword search may miss these because the words differ. Vector search can still connect them through meaning.

Score Fusion

Hybrid search needs a way to combine keyword scores and vector scores. This is called fusion.

Two common fusion approaches are:

The goal is to produce one ranking that respects both exact matches and semantic similarity.

Alpha Weighting

Many hybrid search systems expose a parameter that controls the balance between keyword search and vector search. This is often called alpha or weight.

The right balance depends on the domain. Technical documentation may need more keyword precision than a general recommendation system. Customer support may benefit from a balanced approach.

Hybrid Search With Filters

Hybrid search can also use metadata filters. This is important because even a strong hybrid ranking still needs structured constraints.

Filters decide what is eligible. Hybrid search decides how eligible results should be ranked using both lexical and semantic signals.

Why Hybrid Search Helps RAG

RAG systems need reliable retrieval. If retrieval misses the right context, the language model may produce weak or incorrect answers.

Hybrid search helps RAG because user questions often contain both exact terms and broad intent. A good retriever should handle both.

- Vector search helps retrieve conceptually related chunks.
- Keyword search helps preserve exact terminology.
- Filters keep retrieval inside the right tenant, role, status, or source.
- Fusion creates one ranked list for the prompt or reranker.

This makes hybrid search a strong default for knowledge bases, support assistants, technical documentation, policy search, and enterprise RAG systems.

When Hybrid Search Works Best

Hybrid search is especially useful when the content has a mix of natural language and exact terms.

Implementation Example: Weaviate

Weaviate is a useful implementation example because its hybrid search combines BM25 keyword search with vector search and fuses the result scores. It also supports filters and alpha weighting.

In this example, `alpha=0.5` gives keyword and vector search an even balance. The filters keep retrieval inside published content for one tenant. The returned metadata helps inspect how the final score was built.

Changing the alpha value changes the balance. A lower value leans more toward keyword relevance. A higher value leans more toward vector similarity.

Best Practices

1. Use hybrid search when exact terms and semantic meaning both matter.
2. Tune the keyword/vector balance with real queries, not only demo examples.
3. Use filters for tenant, permission, status, product, source, and freshness constraints.
4. Inspect score explanations when results look surprising.
5. Evaluate hybrid search with queries that include acronyms, codes, names, and paraphrases.

6. Consider a reranker after hybrid retrieval when answer quality depends on precise ordering.

7. Measure retrieval quality for both keyword-heavy and meaning-heavy queries.

Summary

Hybrid search combines keyword search and vector search into one retrieval method. Keyword search preserves exact term matching. Vector search captures semantic meaning. Fusion combines both signals into one ranked result list.

For many production systems, hybrid search is stronger than either method alone. It is especially useful for RAG, technical documentation, customer support, product search, and enterprise search because it can match exact terms while still understanding the user's intent.